به نام خدا

تمرین سوم درس داده‌کاوی

ترم بهار ۱۴۰۲

توضیحات تمرین:

- پاسخ به این تمرین به صورت انفرادی می‌باشد.
- لطفا سوالات را به ترتیب پاسخ دهید.
- در صورت ابهام درباره‌ی تمرین با ایمیل درس با تدریسیاران در ارتباط باشید.

dm.spring1402@gmail.com

- مهلت ارسال تمرین تا **ساعت ۲۱:۵۹ دقیقه روز شنبه ۲۰ خرداد** می‌باشد.
- تمرین شامل دو بخش **تئوری و عملی** می‌باشد.
- فایل‌های ارسالی شما باید یک فایل pdf گزارش (شامل جواب سوالات تئوری و سوالات بخش عملی)، و همچنین شامل کدهای شما باشد، که لطفا آنها را تحت یک فایل zip بارگزاری نمایید.
- فرمت فایل zip شما باید به شکل زیر باشد:

HW3-[student_number].zip (برای مثال HW3-9831011.zip)

فهرست

بخش تئوری

سوال اول

یک مجموعه داده از حیوانات مختلف به همراه ویژگی‌هایشان را در اختیار داریم، می‌خواهیم با استفاده از روش‌های خوشه‌بندی میزان شباهت هر دو حیوان به هم را از ۱ (کمترین) تا ۳ (بیشترین) مشخص نماییم. برای مثال میزان شباهت شیر و پلنگ ۳ و میزان شباهت شیر و گوسفند ۱ می‌تواند باشد. الگوریتمی ارائه دهید که این امر را به صورت غیرنظارت‌شده[1] ممکن سازد.

سوال دوم

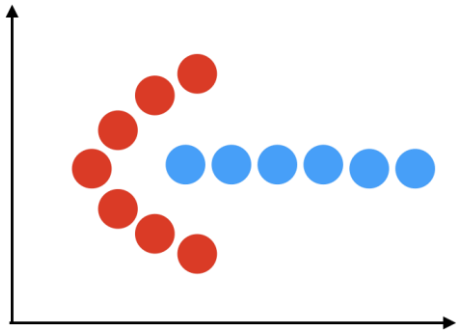
می‌دانیم که در الگوریتم خوشه‌بندی برای تابع مجاورت[2] موارد مختلفی را می‌توان استفاده کرد، در موارد زیر اثبات نمایید که نقطه نهایی که به عنوان مرکز انتخاب می‌شود چه نقطه‌ای است. (در رابطه زیر D مجموعه تمامی نقاط داده و C مجموعه تمامی مراکز خوشه ها می‌باشد.)

$$\sum_{d\in D}\sum_{c\in C} f(d,c)$$

- نرم۱ $(f(d,c) = |d-c|)$
- نرم۲ $(f(d,c) = \|d-c\|_2^2)$

سوال سوم

الف) فرض کنید داده‌های زیر را می‌خواهیم به ۲دسته مختلف دسته‌بندی کنیم، پیش‌بینی شما از اجرا الگوریتم k-means را از داده‌های زیر بیان کنید و علت این پیش‌بینی را هم ذکر نمایید.



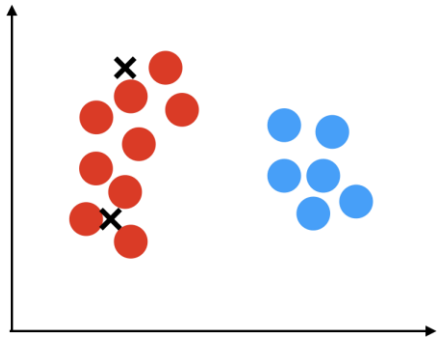
ب) آیا استفاده از روش DBSCAN میتواند برای داده‌های بالا عملکرد بهتری داشته باشد؟ علت را توضیح دهید.

ج) توضیح دهید در چه زمانی خوشه‌بندی بر مبنای چگالی عملکرد مناسبی نخواهد داشت؟ مثال بزنید.

[1] Unsupervised
[2] Proximity Function

سوال چهارم

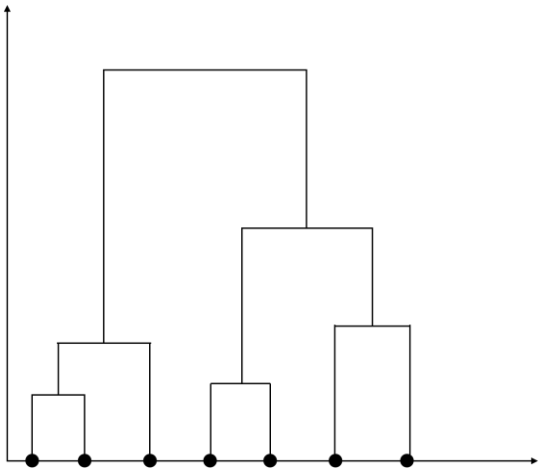
الف) نتیجه اعمال الگوریتم k-means را بر روی داده‌های زیر مشخص کنید. (ضرب در بیانگر مراکز اولیه است)



ب) برای حل مشکل بالا از راهکارهای گوناگونی استفاده می‌شود در رابطه با هر یک از این راهکارها را تحقیق کرده و مزایا و معایب آن‌ها را توضیح دهید

- استفاده از medoid به جای median
- انتخاب نقاط اولیه به شکلی که بیشترین فاصله را از هم داشته باشند
- انتخاب نقاط اولیه بر اساس توزیع داده‌ها
- انتخاب چندباره مراکز اولیه برای رسیدن به جواب مناسب

ج) دندروگرام زیر، انجام خوشه‌بندی سلسله مراتبی را بر روی یک مجموعه‌دادگان را نشان می‌دهد، با توجه به دندروگرام مشخص نمایید که اگر بخواهیم بر روی داده‌های زیر الگوریتم k-means را اجرا نماییم بهتر است که چه مقداری را به k دهیم.



سوال پنجم

ماتریس زیر را در نظر بگیرید با استفاده از روش PCA داده‌ها را به یک بعد انتقال داده و ماتریس داده حاصل را بدست آورید.

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ -1 & -1 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}$$

سوال ششم

با فرض آستانه پشتیبانی[3] برابر 0.3 و آستانه اطمینان[4] برابر 0.4 ، مجموعه تمام قوانین انجمنی ممکن را بنویسید

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke
6	Milk, Diaper, Coke
7	Bread, Diaper, Beer

سوال هفتم

با فرض آستانه پشتیبانی $\frac{1}{3}$ و آستانه اطمینان $\frac{2}{3}$، مجموعه آیتم‌های پرتکرار را به دست آورید. در مرحله بعد مجموعه‌ی تمام قوانین انجمنی ممکن را به دست آورید.

تراکنش	آیتم‌ها
T1	$\{a, b, c\}$
T2	$\{d, c\}$
T3	$\{a, b, c\}$
T4	$\{d, b\}$
T5	$\{a, e\}$
T6	$\{a, e, d\}$
T7	$\{b, c\}$
T8	$\{a, b, c, d\}$

[3] Support
[4] Confidence

بخش پیاده‌سازی

در این برنامه قصد داریم به شکل عملی با خوشه‌بندها آشنا بشویم و پیاده‌سازی از خوشه‌بندها را به کمک زبان پایتون انجام بدهیم. برای شروع کار دیتاستی در اختیار شما قرار گرفته است. این دیتاست مجموعه‌ای است از نقاط که به شکل تصادفی در صفحه قرار گرفته‌اند.

در مرحله بعد لازم است که ماتریس شباهت دادگان را به دست آورید؛ به این صورت که برای هر یک از داده‌های درون مجموعه داده، برداری از ویژگی‌هایش در نظر بگیرید و ماتریس شباهت را برای مجموعه دادگان به دست آورید. برای معیار شباهت نیز، یک بار به کمک معیار شباهت کسینوسی[5] و یکبار به کمک فاصله اقلیدسی[6] ماتریس شباهت را به دست آورید.

بعد از تشکیل ماتریس شباهت، به مرحله‌ی پیاده‌سازی الگوریتم خوشه‌بندی خود می‌رسید. در این مرحله تعداد خوشه ها را برابر ۴ در نظر بگیرید و مراکز اولیه خوشه را به شکل تصادفی انتخاب نمایید. در نهایت الگوریتم خوشه‌بندی k-means را برای خوشه‌های اولیه خود پیاده سازی کنید و خوشه‌بندی را تا نقطه‌ای که به دقت مناسبی برسید ادامه دهید و خوشه‌های خود را تشکیل دهید.

در آخر لازم است که نموداری براساس خوشه‌های تشکیل شده نمایش دهید و عکس نتایج و نمودار (برای هر دو حالت معیار شباهت) را به همراه کد خود ارسال نمایید. توصیه می‌شود برای این کار از یک Jupyter Notebook استفاده شود.

[5] Cosine Similarity
[6] Euclidean Distance

شبه‌کد پیاده‌سازی الگوریتم k-means:

```
# Function: K-Means
# -------------
# K-Means is an algorithm that takes in a dataset and a constant
# k and returns k centroids (which define clusters of data in the
# dataset which are similar to one another).
def kmeans(dataSet, k):

    # Initialize centroids randomly
    numFeatures = dataSet.getNumFeatures()
    centroids = getRandomCentroids(numFeatures, k)

    # Initialize book keeping vars.
    iterations = 0
    oldCentroids = None

    # Run the main k-means algorithm
    while not shouldStop(oldCentroids, centroids, iterations):
        # Save old centroids for convergence test. Book keeping.
        oldCentroids = centroids
        iterations += 1

        # Assign labels to each datapoint based on centroids
        labels = getLabels(dataSet, centroids)

        # Assign centroids based on datapoint labels
        centroids = getCentroids(dataSet, labels, k)

    # We can get the labels too by calling getLabels(dataSet, centroids)
    return centroids
```